

# Putting Recommendations on the Map – Visualizing Clusters and Relations

Emden Gansner, Yifan Hu, Stephen Kobourov, Chris Volinsky  
AT&T Labs – Research, 180 Park Ave, Florham Park, NJ 07932  
{erg, yifanhu, skobourov, volinsky}@research.att.com

## ABSTRACT

For users, recommendations can sometimes seem odd or counterintuitive. Visualizing recommendations can remove some of this mystery, showing how a recommendation is grouped with other choices. A drawing can also lead a user's eye to other options. Traditional 2D-embeddings of points can be used to create a basic layout, but these methods, by themselves, do not illustrate clusters and neighborhoods very well. In this paper, we propose the use of geographic maps to enhance the definition of clusters and neighborhoods, and consider the effectiveness of this approach in visualizing similarities and recommendations arising from TV shows and music selections. All the maps referenced in this paper can be found in [www.research.att.com/~volinsky/maps](http://www.research.att.com/~volinsky/maps).

## 1. INTRODUCTION

Information visualization techniques are often essential in helping to make sense out of large data sets. High-dimensional data can be visualized as a collection of points in 2-dimensional space using principal component analysis [12], multidimensional scaling [15], force directed algorithms [6], or non-linear dimensionality reduction like LLE/Isomap [20, 22]. These embedding algorithms tend to put similar items next to each other. Visual examination often suffices to identify the presence of clusters. Sometimes, however, the clusters are not as easy to see and additional visual clues are needed to highlight them. One possibility is to use cluster analysis algorithms, such as  $k$ -means or hierarchical clustering algorithms [11, 16] to explicitly define clusters. The points and labels can then be colored based on the clustering. While in small examples it is possible to convey the cluster information just with the use of colors and proximity, this becomes difficult to do with large data. Common problems include dense clusters where labels overlap each other and clusters that lack clearly defined boundaries.

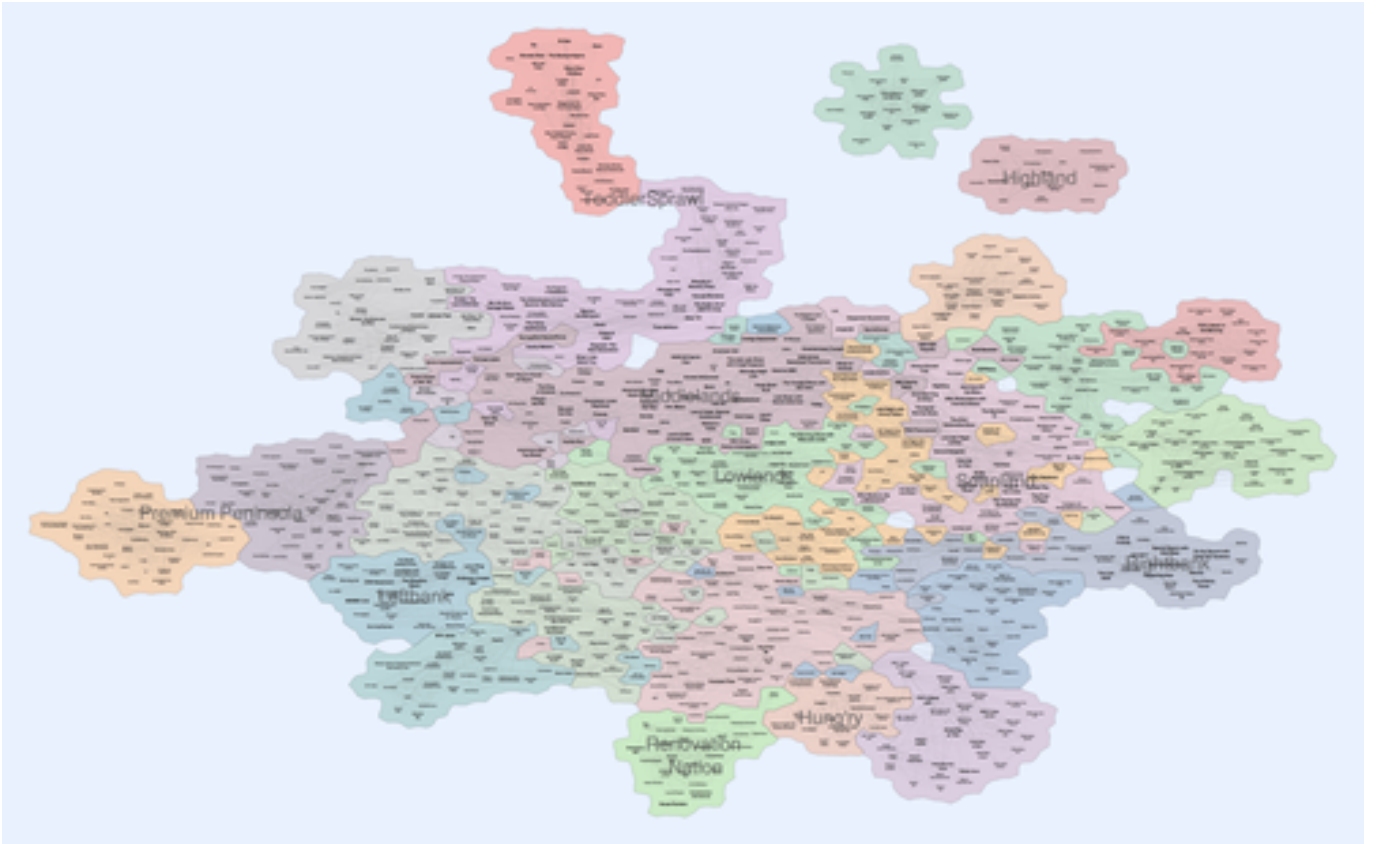
In this paper we propose the use of maps as a way to achieve this explicit visual definition of clusters. There are several reasons that such a representation can be more useful. First, by explicitly defining the boundary of the clusters and coloring the regions, we make the clustering information clear. Second, as most dimensionality reduction techniques lead to a 2-dimensional positioning of the data points, a map is a natural generalization. Finally, while graphs, charts, and tables often require considerable effort to comprehend,

a map representation is more intuitive, as most people are very familiar with maps and even enjoy carefully examining maps. Applying this approach to a data set containing show-show similarities between 1000 TV programs results in the map in Figure 1, which conveys clustering information much better.<sup>1</sup>

We have considered several different approaches for generating maps, depending on (1) how we obtain the 2D positions for the data points, (2) how we cluster the data points, and (3) how we represent the resulting layout and clustering as a map. Depending on the application, some choices are more suitable than others. Consider the case when the data points are TV programs and the goal is to visualize “TV-Land”. In this case, it is highly desirable to obtain a map in which similar programs are placed close to each other and, even better, grouped in *countries* such as *Sportsitania*, *Newsistan*, *ToddlerSprawl*, etc. Once the data is represented as countries on a map, recommendations can be visualized statically and interactively.

The map metaphor becomes more powerful as a user becomes familiar with the canonical map layout. Humans are comfortable with map-related concepts: items within a country are similar to each other; areas separated by a mountain range are difficult to connect; islands might have atypical qualities, etc. We hope through this work to extend the familiar map metaphor to the world of recommendations. Thus, by putting items like TV shows on a map, we can borrow map-related cognitive concepts. In a user-driven mode, a personalized “heat map” is generated, where regions of low interest are colored with cool colors and regions containing highly recommended shows are colored with a hot color. These maps are generated dynamically based on user preference and the available TV shows at this moment in time. The user can interactively explore the map to find related shows in the same region, or in neighboring regions. Building “roads” between regions, visiting “islands” and “traveling cross-country” are all metaphors that can have meaning in the recommendation space. In a user-passive mode, recommendations are highlighted on the map. These maps help the user (as well the designers of the recommender system) understand why certain shows are recommended. By highlighting both the recommended shows and shows the user has watched on the same map, the user understands the proximity of the recommended items and watched items, thus the rationale for the recommendation.

<sup>1</sup> This paper contains zoomable high resolution images; all the images are also available at [www.research.att.com/~volinsky/maps](http://www.research.att.com/~volinsky/maps).



**Figure 1:** TVLand: a map based view of the TV similarity graph, depicting the relationship among the 1000 most watched TV shows. Each show is linked to 10 most similar shows and map clusters, or *countries* are represented by different colors. Zoom into the graph to read the labels and see more structure or view the images at [www.research.att.com/~volinsky/maps](http://www.research.att.com/~volinsky/maps).

## 2. THE MAPPING ALGORITHM

The input to our algorithm is a relational data set from which we extract a graph  $G = (V, E)$  with  $|V| = n$  and  $|E| = m$ . The set of vertices  $V$  corresponds to the objects in the data, e.g., TV programs, and the set of edges  $E$  corresponds to the relationship between pairs of objects, e.g., the similarity between a pair of shows. In its full generality, the graph is vertex-weighted and edge-weighted, with vertex weights corresponding to some notion of the importance of a vertex and edge weights corresponding to some notion of the distance between a pair of vertices. In the case of TV programs, the importance of a vertex can be determined by the popularity of a show as derived from the total number of viewers or by the number of hours watched by one particular user. The weight of an edge connecting two programs can be determined by textual comparison of their program description using a content based model (e.g., [2]), or by analyzing viewership using a collaborative filtering based model [8].

In the next step, the graph is embedded in the plane. Possible embedding algorithms include principal component analysis [12], multidimensional scaling (MDS) [15], force directed algorithm [6], or non-linear dimensionality reduction like LLE/Isomap [20, 22]. A cluster analysis is then performed in order to group vertices into clusters. It is important to match the clustering algorithm to the embedding algorithm. For example, a geometric clustering algorithm

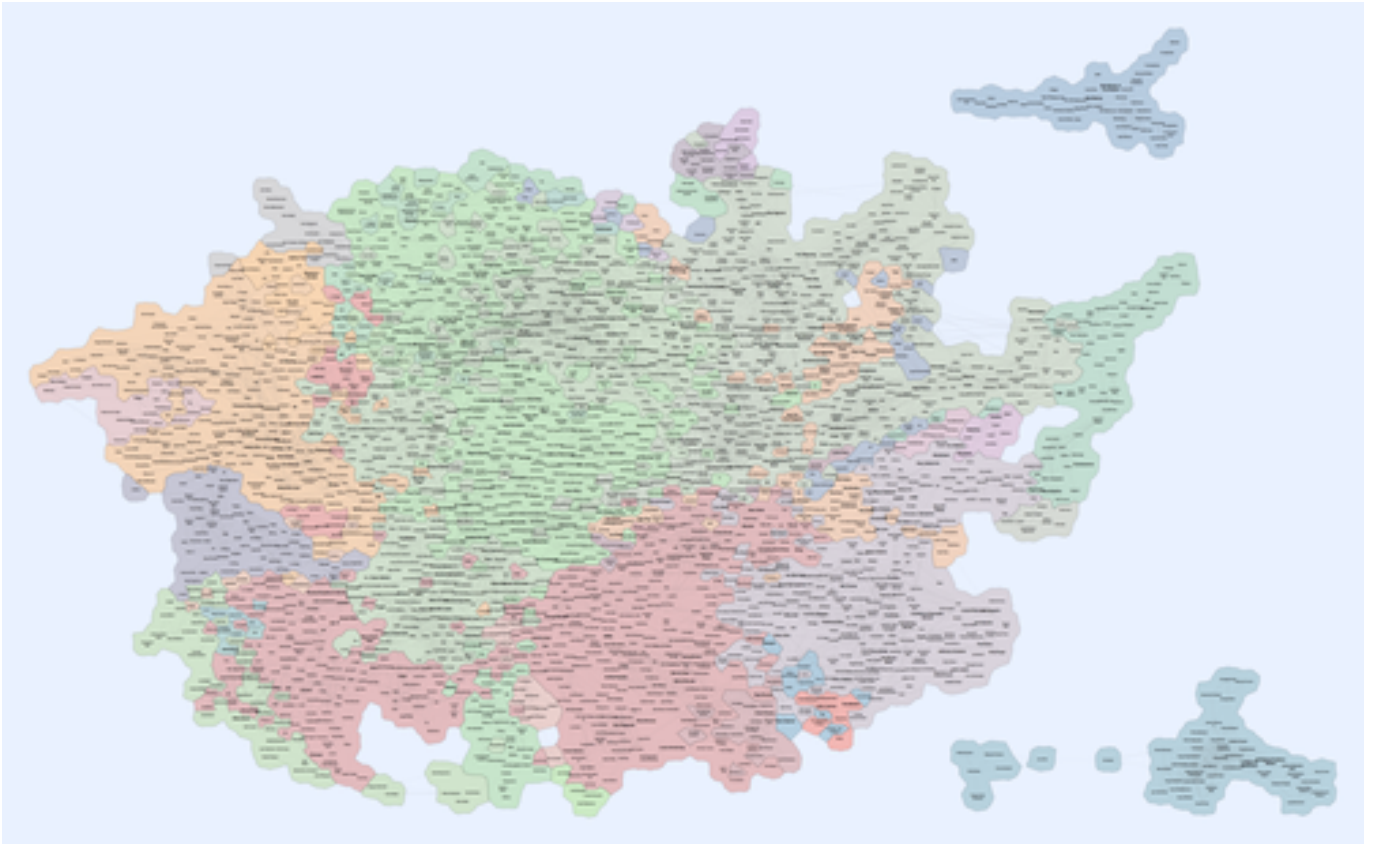
such as  $k$ -means [16] may be suitable for an embedding derived from MDS, as the latter tends to place similar vertices in the same geometric region with good separation between clusters. On the other hand, with an embedding derived from a force directed algorithm [6], a modularity based clustering algorithm [17] is appropriate. The two algorithms are strongly related, as pointed out in the recent findings by Noack *et al.* [18], and therefore we can expect vertices that are in the same cluster to also be geometrically close to each other in the embedding.

The final map is obtained in the last step. Using the layout information, a Voronoi diagram of the vertices is created. A naive approach would be to form the Voronoi diagram of the vertices, together with four points on the four corners of the bounding box; see Fig. 2(a). This would result in aesthetically unappealing maps with unnatural outer boundaries and sharp corners. A more natural appearance can be obtained by placing some random points that are sufficiently far away from the set of real vertices, which leads to more rounded boundaries. The randomness of the points on the outskirts also gives rise to some randomness of the outer boundaries, thus making them more realistic and natural; see Fig. 2(b).

In most of our maps *countries* have different desirable sizes (e.g., in TVLand the size is determined by a show's popularity). We accomplish this by associating larger *countries*







**Figure 8:** Map of MusicLand

and *CSI*, talk shows such as *Late Show with David Letterman* and *Tonight Show with Jay Leno*, entertainment shows such as *TMZ* and *Access Hollywood*, talent shows such as *American Idol* and sports programming such as *NASCAR Sprint Cup* and *NBA Basketball*. Directly below the Middlelands of popular shows lie the Lowlands of reality-based court shows *Judge Judy*, *Judge Joe Brown* and *The People's Court*. In the same cluster we find notorious tabloid talk shows like *Jerry Springer* and *Maury*. Next to the Lowlands is the domain of daytime soap operas with the daytime staples of each network represented; see Fig. 7. ABC's *General Hospital*, *One Life to Live*, *All My Children* share the same space with CBS's *Young and the Restless*, *Guiding Light*, *As the World Turns* and *The Bold and the Beautiful*. NBC's contribution to this cluster is *Days of Our Lives*.

### *ToddlerSprawl.*

The NW corner of TVLand consists almost exclusively of kid-oriented shows. The northernmost point has a prominent peninsula dominated by Nickelodeon powerhouses like *Dora*, *Ni Hao* and *Go*. This connects to the mainland via a Disney cluster with crowd-pleasers *Hannah Monana* and *The Suite Life of Zack and Cody*. ToddlerSprawl continues on the north shore of the mainland with cross-over hit shows like *SpongeBob SquarePants*. At the NW tip of ToddlerSprawl are Cartoon Network's *Johnny Test* and mainstays *Tom and Jerry*, *Flintstones* and *Scooby Doo*. Conspicuously absent are *Sesame Street* and *Barney and Friends* which were clustered separately with other PBS shows.

### *Premium Peninsula.*

The western coast of TVLand consists of two monolithic clusters corresponding to the premium channels HBO and Showtime. The HBO peninsula is represented by popular original series such as *Real Time with Bill Maher* and *Big Love* and several blockbuster movies. Similarly, its neighbor, the inland Showtime cluster has some original series such as *The L Word* and dozens of popular movies. Most of the other movies in TVLand can be found in the immediate vicinity of HBO and Showtime.

## 3.2 The land of music

We now turn to the visualization of the land of music. We collected data from a web crawl of the `last.fm` website. As an Internet radio and music community website, it has over 30 million users. Using a music recommender system, `last.fm` recommends music based on user profiles. Over several years the recommender system has collected information about how one musician is related to another in terms of how many listeners of one also enjoy the other. For each musician, the `last.fm` website lists the top 250 related musicians. For example, Beethoven is considered to have “super similarity” to Mozart, Bach, Brahms etc., “very high similarity” to Mendelssohn, Schumann, Vivaldi, etc.<sup>3</sup> The website also provides the number of listeners of each musician. In April 2009, we crawled the website by starting

<sup>3</sup>Note that we do not know `last.fm`'s formal definition of “super similarity”, “very high similarity”, or “high similarity”.



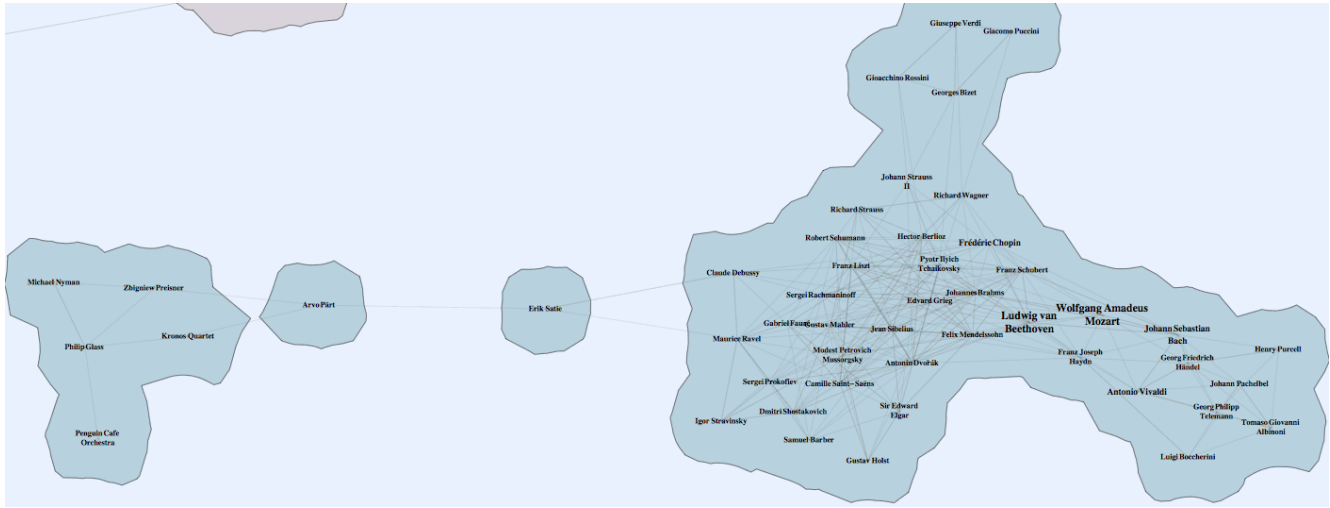


Figure 9: Classical archipelago

with Beethoven, and the top 20 musicians most similar to Beethoven, provided that each has at least 100,000 listeners. We then found the top 20 most similar musicians to each of those with at least 100,000 listeners and proceed recursively. Our crawl yielded a graph with 2782 musicians, with edge weights the similarity between musicians. We further prune this graph by only taking edges that have “super similarity” and after taking the non-trivial components we end up with 2588 vertices. We then lay out the graph, cluster the vertices, and generate the MusicLand map; see Fig. 8.

### The Mainland.

The vast majority of musicians and bands is located in the continent. While it is not as easy to spot major trends along the main axes, many of the clusters are well-defined and neighboring clusters make sense from a musical point of view. The *Rocky Coast* cluster on the east shore begins with *Eric Clapton* and the *Beach Boys* in the south, goes through the *Rolling Stones* and *Neil Young* in the middle before reaching the *Metallands* cluster of *AC/DC* and *Black Sabbath* in the north. In close proximity to the southern edge of the Rock cluster is a small but major *Beatles* cluster. To the north is the *Grungelands* cluster with *Nirvana*, *Pearl Jam*, and *Soundgarden*.

On the west coast of the main continent we find the *Electrolands* cluster of lounge and acid jazz, anchored by *Thievery Corporation*, *Morcheeba*, and *Massive Attack*. To the north are avant-garde electronic bands like *Autechre*, *Aphex Twin* and *Boards of Canada*. In the extreme west are electronic classics *Tangerine Dream*, *Vangelis*, and *Jean-Michelle Jarre*. To the south is a compact and well-defined cluster of dance music represented by *BT*, *Paul Oakenfold*, and *ATB*.

In the center of the map, there is a large number of indy bands, the *Indyana* cluster, dominated by *The Decemberists*, *The Shins*, and *Wilco*. Nearby is a well-defined concentration of female singer-songwriters: *Alanis Morissette*, *Tory Amos*, *Sarah McLachlan*, and *Tracy Chapman*.

### The Islands.

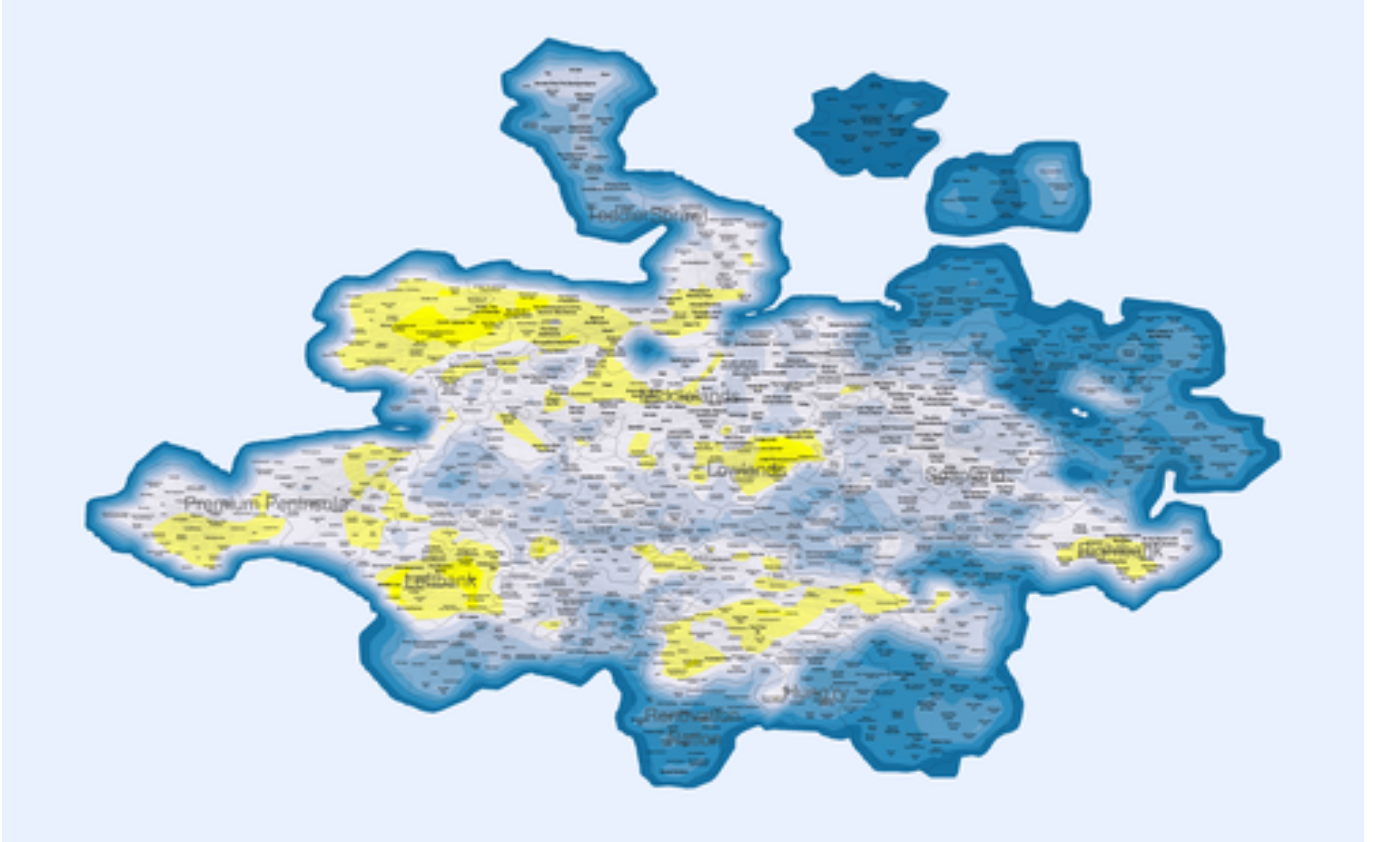
There are two notable islands regions in MusicLand: in the NE is Reggae island, while the chain of islands off *Rocky Coast* in the NW make up the *Classical Archipelago*. It is

easy to find some general patterns in the layout of *Classical archipelago*; see Fig. 9. Two cardinal directions are W-E and N-S. Along the first axis, the west corner contains modern composers such as *Ravel*, *Satie*, and *Pärt*, while the east corner contains 17th century composers such as *Bach*, *Handel*, and *Albinoni*. Along the second axis, the north corner has a high concentration of opera composers such as *Verdi*, *Rossini*, and *Puccini*, whereas the south corner has more orchestral and instrumental composers such as *Holst*, *Elgar*, and *Stravinsky*. Not surprisingly, *Mozart* and *Beethoven* are the most popular composers in the classical music cluster. The islands of *Erik Satie* and *Arvo Pärt* connect the big island in the east with the west-most island of contemporary classical music represented by minimalists *Philip Glass* and *Michael Nyman*.

## 4. VISUALIZING RECOMMENDATIONS

Once a global map of the item-item similarity is available, a personalized recommendation “heat map” can be generated, where regions of low interest are colored with cool colors, and regions containing highly recommended shows are colored with a hot color. Figure 10 shows such a heat map, where shows are scored using a factorization based recommender [10], with blue color for shows that score low, and yellow for shows that score high. We generate such maps dynamically based on the viewing preference of the user, and based on what TV shows are available at this moment in time, much like a personalized weather forecast, but for TV shows. These maps uniquely capture the viewing preferences of the user/household, and evolve as the availability of TV shows, and as the user’s taste, change with time. We can also generate a heat map profile, determined by how often the user watches certain shows over a fixed time period, say, a week or a month.

Using the map interface, recommendations can be made in several different ways. In a user-driven mode, the heat map is presented to the user for exploration. The user sees the areas of concentration of recommended shows or their viewing behavior, and can explore the map to find other shows in the same countries, or in neighboring countries. In some cases, a user might want to explore faraway lands for something new, or complete a cross-country trip over time. In a user-passive

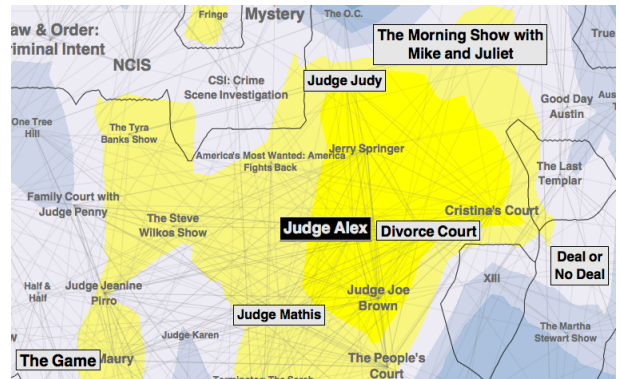


**Figure 10:** Personalized recommendation heat map for a typical but fictitious user. Regions where the highest recommended shows sit are colored yellow. Regions of low scores are colored with dark cool colors.

mode, recommendations are made by the recommender system and are highlighted on the map, along with the shows that were most similar. In this mode, the map allows the user to understand why the recommendations are made. For example, Figure 11 shows that *Judge Alex* is recommended because the user watched *Divorce Court*, *Judge Judy*, and *Judge Mathis*. In addition to the recommended show, the map shows other nearby shows that are related and which might be worth exploring.

Note that the “hot” areas on the map in Figure 10, include news shows and kids shows. This could be an indication that there are children and adults that household. Our TV data is on the household level, and a household may contain various individuals with widely varying preferences. The visualization of household viewing behavior on a map could possibly separate out different elements of the household and allow each member to get more personalized recommendations, e.g., the 8-year-old boy will look for interesting shows in *ToddlerSprawl*, while the teenagers look in the *Premium Peninsula* and the parents in *Newsistan*.

While *last.fm* does not explicitly make recommendations, it does so implicitly, by playing music that is considered related to that often requested by the user. For example, one of the authors of this paper is a fan of the music groups *Thievery Corporation*, *Tosca*, *Kruder & Dorfmeister*, and *Fila Brazillia*. After several months of frequently requesting that *last.fm* play music similar to the above four groups, a look at the list of music most often played reveals that all



**Figure 11:** Explaining a recommendation. Framed labels with a light background are shows the user watched, with font size proportional to the amount of time spent. The recommended show is framed with a black background.

bands in the immediate vicinity (such as *Sofa Surfers*, *Air*, *Waldeck*) were played with high frequency; see Fig. 12.

We envisage our maps as a navigational interface to interactive media services such as video/movie/music-on-demand, in which each title is a live link to the video/audio recording, and the user can zoom in and out to explore the land of TV/movies/music, much as online maps are used, except

